Chance and Consensus in Peer Review

Stephen Cole, Jonathan R. Cole, Gary A. Simon

Summary. An experiment in which 150 proposals submitted to the National Science Foundation were evaluated independently by a new set of reviewers indicates that getting a research grant depends to a significant extent on chance. The degree of disagreement within the population of eligible reviewers is such that whether or not a proposal is funded depends in a large proportion of cases upon which reviewers happen to be selected for it. No evidence of systematic bias in the selection of NSF reviewers was found.

The National Science Foundation (NSF) employs one form of the peer review system in making research grants. For each application for a grant, an NSF program director selects a group of scientists, generally four or five, who are knowledgeable in the relevant subject matter, to act as referees. Each reviewer is sent a copy of the proposal and asked to evaluate it on the basis of its scientific merit and the ability of the principal investigator. Ability of the principal investigator is generally defined as the quality of his or her recent scientific performance. Each reviewer is asked to make substantive comments and to assign one of five ratings to the proposal: excellent, very good, good, fair, or poor. We ask whether the procedure employed by NSF is an equitable and a rational one.

For the past 5 years, as consultants to the National Academy of Sciences' Committee on Science and Public Policy (COSPUP), we have been conducting a study of NSF's peer review (*1*). This work has been divided into two phases. In this article we report on the second phase of that extended inquiry. Since they represent the point of departure for the experiment described here, we recapitulate briefly the principal results of the first phase, which were based on 75 extended interviews with NSF staff, on analysis of 1200 proposals drawn from ten NSF programs, and on the substantive comments of reviewers of 250 of these proposals.

1) There is a high correlation between reviewer ratings and grants made. If one attaches numerical values to the ratings, say from 10 for poor to 50 for excellent, the mean scores predict with a high degree of accuracy which proposals will be funded and which will be denied. Whether or not NSF program directors actually compute statistical averages from the ratings and use them in decision-making, the statistical average of the ratings turned out to be highly correlated with the actual decision rules employed by the program directors.

2) For the 1200 proposals there was not a high correlation between grants awarded and measures of the previous scientific performance of the applicants. This result was unexpected, since one of the stated evaluation criteria is the ability of the applicants to conduct the research proposed.

3) Reviewers at major institutions did not treat proposals from applicants at major institutions more favorably than did reviewers from lesser institutions. In fact, there was a tendency in the opposite direction.

4) Professional age (length of career) had no strong effect on either ratings received or the probability of receiving a grant.

5) There were low or moderate correlations between reviewer ratings (and the funding decision) and the following characteristics of the applicants: prestige rank of current academic department, academic rank, geographic location, NSF funding history over the previous 5 years, and locus of Ph.D. training (*2*).

Because proposals from eminent scientists do not have substantially higher probabilities of receiving favorable ratings than proposals from scientists who are not eminent, we concluded that the peer review system employed by NSF was essentially free of systematic bias. We now want to take up the further question of whether the system as currently employed is, in addition to being equitable, a rational one. In particular, we are concerned with the role of chance in obtaining an NSF grant. A rational system would minimize random elements and maximize the influence of both the quality of the proposal and the ability of the principal investigator to perform the research.

The COSPUP Experiment

The second phase of the study was designed to tell us, among other things, whether or not the program directors were predetermining funding decisions by their selection of reviewers—that is, whether independently selected panels of reviewers would reach similar conclusions.

In the spring of 1977, the NSF provided us with 150 proposals—50 each from the programs in chemical dynamics, economics, and solid-state physics—upon which decisions had been made recently; half the proposals in each program had been funded and half had been declined. We then obtained other reviewers for those proposals. In order to select the new reviewers we utilized a panel of 10 to 18 experts in each of these fields, most of them members of the National Academy of Sciences (*3*). Each proposal was sent to two members of this panel, each of whom selected six or more reviewers for it. This gave us a list of approximately 12 reviewers for each proposal.

Some have argued that the highly specialized state of modern science would not permit more than a dozen or so scientists to be capable of reviewing any given proposal. The COSPUP experiment enabled us to test this hypothesis. If the number of eligible reviewers was, in fact, small, we would expect that a fairly high proportion of the original NSF reviewers would also have been selected by the experimental selectors. In each of the three programs, about 80 percent of the NSF reviewers were not selected by either of the two COSPUP selectors, about 15 percent were selected by one of them, and about 5 percent were selected by both. These data suggest that the pool

Stephen Cole is professor of sociology at the State University of New York, Stony Brook 11794. Jonathan R. Cole is professor of sociology and director of the Center for the Social Sciences at Columbia University, New York 10027. Gary A. Simon is associate professor of applied mathematics at the State University of New York, Stony Brook 11794.

0036-8075/81/1120-0881$01.00/0 Copyright © 1981 AAAS

Table 1. Correlation of mean ratings of NSF reviewers and COSPUP reviewers on grant applications (N = 50) in each of three NSF programs. Numerical values were assigned to the original ratings as follows: excellent = 50, very good = 40, good = 30, fair = 20, and poor = 10. Figures in parentheses are standard deviations.

Field	Mean ratings		Correlation coefficient
	NSF	COSPUP	
Chemical dynamics	37.7 (±5.85)	35.0 (±6.45)	.595
Economics	33.6 (±9.76)	31.5 (±9.28)	.659
Solid-state physics	38.2 (±6.15)	35.5 (±6.33)	.623

of eligible reviewers for most proposals is at least of size 10 and, given the low overlap rates we found, we would predict that if other equally qualified selectors were employed we would find it to be substantially larger than 20 (*4*). Of course, in actual practice there is not a clear distinction between eligibles and noneligibles, and the numbers certainly vary according to subspecialties. Since the pool of eligible reviewers for most proposals is substantially larger than the actual number of reviewers used by NSF, and since there was little overlap between NSF and COSPUP reviewers, we want to consider the extent to which the two sets of reviewers agreed upon the merits of the proposals.

In general, the COSPUP reviewers tended to give slightly lower scores than did the NSF reviewers (Table 1). For example, for the 50 chemical dynamics proposals, the grand mean of the NSF reviewers' ratings was about 38 on the 10-to-50 scale, that of the COSPUP reviewers about 35. The experimental reviewers may have been slightly harsher in their evaluations because they knew that their ratings would have no effect on the careers of the applicants. The correlations between the mean NSF rating and the mean COSPUP rating for each proposal (Table 1) are moderately high (.60, .66, and .62). Proposals that are rated high by NSF reviewers tend also to be rated high by the independent sample of reviewers used by COSPUP. The match is, however, less than perfect.

The findings presented thus far do not address one of the fundamental questions for evaluating the peer review system at NSF: How many funding decisions would be reversed if they were determined by the COSPUP ratings rather than by the procedures followed by NSF?

In Fig. 1 we show the rank order of the proposals in each program according to the NSF mean ratings and the mean ratings of the COSPUP reviewers. (Half-integer ranks are the result of ties.) Since the mean ratings generally determine which proposals are funded, and half were funded and half declined, decisions on proposals which were ranked in one half of the range of scores by NSF reviewers and in the other half by COSPUP reviewers would have been reversed by the COSPUP ratings. There were differences of that degree in the ratings of approximately one-quarter of the proposals (Fig. 1).

Reversals

The NSF is faced, of course, with a zero-one decision rule: to fund or not to fund a proposal (*5*). It follows that proposals with mean rankings that are fairly close, or virtually identical, may fall on opposite sides of the dividing line. Therefore, it was almost inevitable that we would find some reversals.

In determining what should be classified as a reversal we rank-ordered the proposals according to their mean COSPUP ratings and assumed that those with the top 25 scores would be funded and the bottom 25 would be declined. We then compared the COSPUP ratings with both the actual NSF decision and the decision NSF would have made if it had relied solely on mean ratings. The NSF funding decision was highly, though not perfectly, correlated with the mean ratings that NSF reviewers had given the proposals; hence the two comparisons yield approximately the same results (Table 2).

If decisions on the 50 proposals were made by flipping a coin, we would expect to obtain a 50 percent reversal rate, on the average. Correlatively, if the COSPUP reviewers were to rate the 50 proposals in such a way that there was complete agreement with the NSF reviewers on which proposals fell into the top 25 and which into the bottom 25, the reversal rate would be zero. Thus, we would expect to find a reversal rate somewhere between zero and 50 percent. In fact, the reversal rate turns out to be between 24 percent and 30 percent for each of the three programs computed in each of the two different ways (Table 2). That is, on 12 to 15 of the 50 proposals in a program the COSPUP reviews led to a different decision from that of the NSF reviews.

We would expect to find some reversals around the cutting point—for example, to find that a proposal ranked 24th by NSF was ranked 26th or 27th by COSPUP. We want to examine the extent to which reversals were common not only at the midpoint but at a distance from it. This is shown in Table 2 by the reversal rates within quintiles (*6*). In chemical dynamics and solid-state physics we find, as expected, the highest reversal rate in the middle quintile. A 50 percent reversal rate for this quintile would not be surprising. In chemical dynamics it is 60 percent in both comparisons and in solid-state physics 49 percent and 43 percent. In economics, on the other hand, we find higher reversal rates in the second and fourth quintiles than in the third. In all three programs there are more than a few reversals in the first quintile. There are, in fact, proposals that were rated in the top quintile by NSF reviewers that would not have been funded had the decision depended on the appraisals of the COSPUP reviewers.

There are several possible explanations for the reversals. Differences between NSF procedures and COSPUP procedures will be considered first. If the two sets of reviewers used different criteria in appraising proposals, the outcome could have differed significantly, creating reversals—for example, if one group of reviewers based their ratings strictly on evaluations of the proposal and the other primarily on the past track record of the applicant. Since the two groups of reviewers were given identical instructions about the criteria, it is unlikely that there were systematic differences of that kind.

Another possible procedural cause of reversals might obtain if NSF and COSPUP selected different types of reviewers. Reviewer differences rather than proposal differences could then result in reversals. Since a comparison of the characteristics of the two groups of reviewers showed few differences, it is likely that they were drawn from the same population.

Assuming that reversals did not result from the procedures employed in the experiment, we are left with two possible substantive explanations. Reversals could result from bias in the way in which the reviewers were selected by either the NSF program director or the COSPUP experiment. If, for example, the NSF program director purposely selected reviewers who would give unrepresentative negative or positive ratings to a proposal, this could create a reversal.

Second, reversals could have resulted from disagreements among fairly selected reviewers using the same criteria. If there is substantial dissensus in the population of eligible reviewers of a given proposal, then it would be possible for equally qualified and unbiased groups of reviewers using the same criteria to differ in the mean rating.

Consider a hypothetical proposal for which there is a population of approximately 100 eligible reviewers. If all 100 were totally agreed about its merits, then any sample of four or five selected at random from the 100 would agree among themselves, and two independently selected samples would not reach different conclusions. However, if the population of eligible reviewers had substantial disagreement about the proposal, two randomly selected samples could yield different mean ratings possibly leading to different outcomes for the proposal. Our data indicate that the reversals in this experiment were a result of such disagreement.

Table 2. Percentage of NSF outcomes (mean rating of NSF reviewers or actual funding decision) reversed in COSPUP rank-order quintiles and overall. Reversals are shifts from the top 25 positions in the COSPUP rank order to the bottom 25 or vice versa.

NSF outcome	Quintile based on COSPUP rating					Overall (N = 50)
	1	2	3	4	5	
			Chemical dynamics			
Mean rating	26	24	60	20	20	30
Decision	26	24	60	20	20	30
			Economics			
Mean rating	20	45	30	45	0	28
Decision	5	45	28	42	0	24
			Solid-state physics			
Mean rating	23	22	49	34	6	27
Decision	16	24	43	29	11	25

Consensus

In order to determine the extent to which the reversals could be explained by bias or disagreement we used analysis-of-variance techniques. Because we did not want to make the usual statistical assumptions (such as normality) which must be made in a standard two-way analysis of variance, we used a components-of-variance model that did not require some of these assumptions but would be useful in answering the same substantive question.

In order to assess the relative magnitude of contributions of the proposal evaluation method and the reviewer to the variation in ratings, we represent the rating y_{ijk}, given by the kth reviewer under method i to proposal j, by

$$y_{ijk} = a_i + b_j + c_{ij} + e_{ijk}$$

where a_i is the overall average rating by evaluation method i (i = 1 for NSF and i = 2 for COSPUP), b_j is the differential effect of proposal j, c_{ij} measures the extent to which the rating on proposal j depends on the evaluation method, and e_{ijk} is the effect caused by the kth reviewer of proposal j by evaluation method i.

We consider a_i to be a fixed quantity and the remaining terms to be random with means equal to zero. Then we can decompose the variance associated with proposals under evaluation method i into three terms:

$$\mathrm{Var}\ (Y_{ijk}) = \sigma_p^2 + \sigma_I^2 + \sigma_{R,i}^2$$

where $\sigma_p^2 = \mathrm{Var}\ (b_j)$ reflects the intrinsic variability of the proposals: $\sigma_I^2 = \mathrm{Var}\ (c_{ij})$ is the variability associated with the interaction of proposals and evaluation method; and $\sigma_{R,i}^2 = \mathrm{Var}\ (e_{ijk})$ is the reviewer variance for method i.

If σ_p^2 is large relative to σ_I^2, $\sigma_{R,1}^2$, and $\sigma_{R,2}^2$, we interpret this to mean that it is relatively easy to distinguish the proposals independent of the evaluation method. However, if σ_I^2 is of the same order of magnitude as σ_p^2, this would suggest that dependence between proposal and evaluation method is masking some of the intrinsic proposal variability. As a consequence, the proposals would be ranked differently under the two evaluation methods. If, as actually occurs in these data, $\sigma_{R,1}^2$ and $\sigma_{R,2}^2$ dominate σ_I^2 and are of the same magnitude as σ_p^2, then reviewer variability will be so pronounced that two different evaluations will give dissimilar rank orders.

The estimates of σ_p^2, σ_I^2, $\sigma_{R,1}^2$, and $\sigma_{R,2}^2$ are presented in Table 3. The dependent variable for the analysis is the rating given the proposal by a reviewer. If we consider all the variance in an entire set of reviews (for example, all reviews done by both NSF and COSPUP reviewers for the 50 proposals), we want to know the sources of variance. There are four possible sources of variance, two of which turned out to be trivial in this study. Consider these four sources and

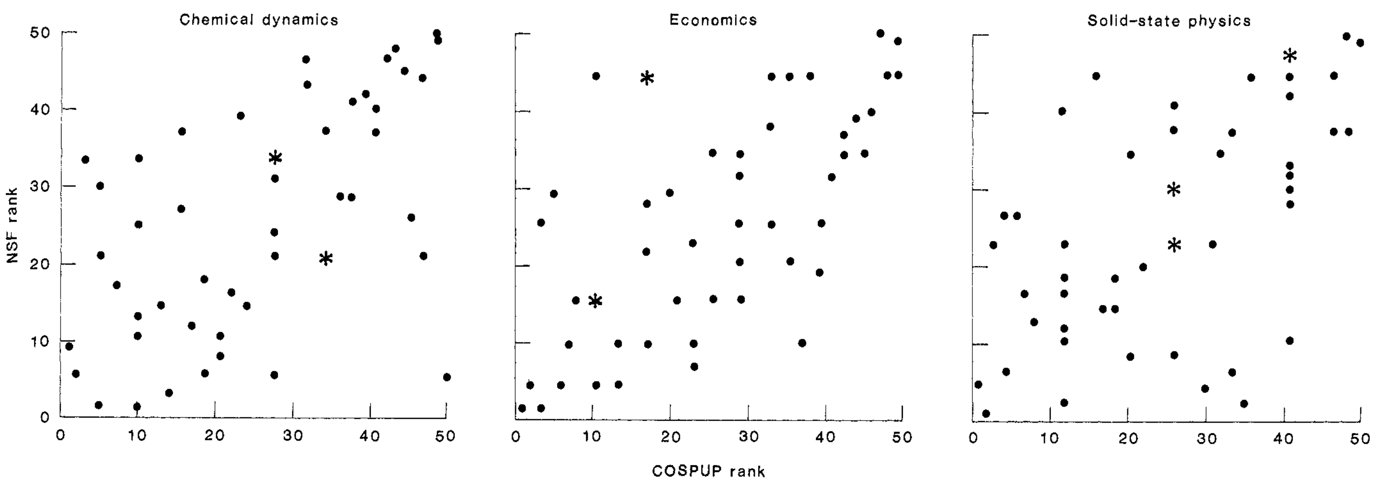

Fig. 1. Rank order of proposals according to mean ratings by NSF and COSPUP reviewers. N = 50 in each program. *Asterisk indicates two proposals with identical ranks.

Table 3. Components of variance of NSF and COSPUP ratings.

Field	Proposal variance $\hat{\sigma}_p^2$	Reviewer variance		Interaction variance $\hat{\sigma}_I^2$	Method difference $\hat{a}_1 - \hat{a}_2$
		NSF $\hat{\sigma}_{R,1}^2$	COSPUP $\hat{\sigma}_{R,2}^2$		
Chemical dynamics	23.67	55.91	56.67	1.18	2.73*
Economics	58.33	89.22	96.25	0.00†	2.14*
Solid-state physics	24.43	48.93	50.24	0.17	2.72*

*NSF higher. †Computed as −1.36.

Table 4. Percentage of total variance in reviewers' ratings accounted for by differences among reviewers of individual proposals. The number in parentheses is the total number of reviewers. For each field there were 50 proposals.

Field	Percent of total variance	
	NSF	COSPUP
Chemical dynamics	60 (242)	53 (213)
Economics	51 (192)	49 (190)
Solid-state physics	43 (163)	47 (182)

the estimated effects for solid-state physics. The results for economics and chemical dynamics have parallel interpretations.

First, reviewers' responses to proposals differ because proposals differ in quality. That is easily dealt with statistically by taking as a rough indicator of the quality of a proposal the mean of all its ratings by both NSF and COSPUP reviewers. This leads to a measure of the variation in quality of proposals (σ_p^2 above) that can be compared with other sources of variation. The estimated proposal variance for the solid-state physics proposals was 24.43.

Second, the NSF review procedures and the COSPUP procedures were not identical. On the average, there may be systematic differences between NSF reviewer responses to all proposals and COSPUP reviewer responses. In fact, this "method effect" can be observed in the differences in the mean ratings of proposals by NSF and COSPUP reviewers. As noted above, the COSPUP reviewers were on average slightly harsher than NSF reviewers. In the NSF-COSPUP comparison the estimated overall difference is 2.72 points, with NSF higher. Since funding decisions are based on rankings, this method effect is not important (but we did not ignore it in the mathematical analysis).

Even after compensating for the average methods effect, reviewers may disagree in their ratings of a proposal because they are members of two groups selected differently—NSF reviewers as opposed to COSPUP reviewers. This "interaction" effect (σ_I^2 above) between proposals and evaluation method is important. It is the key component in estimating whether there appears to be any systematic bias among NSF program directors in the selection of reviewers. If there was bias in the selection of NSF reviewers, or if the two groups of reviewers had significant differences in the way in which they evaluated the proposals *due to any reason*, we would expect the interaction effect to be large. If it is large, then the NSF reviewer group and the COSPUP reviewer group evaluated proposals differently. If it is small, they did not and we would not be able to detect any bias in the selection of the NSF reviewers. It turns out that the estimated interaction σ_I^2 is trifling for each of the three fields, so there is no evidence of disagreement between the two selection methods aside from apparent disagreement resulting from the reviewer variability.

Finally, variation that remains is denoted by $\sigma_{R,i}^2$ above and measures the reviewer variation within a given evaluation method *i*. The reviewer variances were estimated to be 48.93 and 50.24 for solid-state physics. These numbers are rather larger than the estimated proposal variance of 24.43. Thus the reviewer brings to this process a higher variance than does the proposal. Of course, the average of several reviewers will have a lower variance; indeed, the average of four reviewers will have a variance of 48.93/4 = 12.23 (NSF) or 50.24/4 = 12.56 (COSPUP), but these are still not tiny compared to the proposal variance. This hard fact explains why the data exhibit so many reversals; they reflect substantial reviewer variance and not any fundamental disagreement between NSF and COSPUP reviewing methods or substantive evaluations. We may therefore conclude that there was no systematic bias in the way in which NSF reviewers were selected or in the way the two groups of reviewers made their evaluations.

To explain the reversals, then, we must look at two sources of variance: differences among the proposals and differences among the reviewers of a given proposal (Table 3). In the two physical sciences the variance among reviewers of the same proposal is approximately twice as large as the variance among the proposal means; in economics the reviewer variances are about 50 percent larger than the proposal variance. If the pooled proposal mean (the mean of both sets of ratings in each comparison) is taken as a rough indicator of the quality of the application, we can see that the variation in quality among the 50 proposals is small compared to the variation in ratings among reviewers of the same proposal. We have treated the reviewer variances as rough indicators of disagreement among reviewers. In all three fields there is a substantial amount of such disagreement. It is the combination of relatively small differences in proposal means and relatively large reviewer variation that creates the conditions for reversals (*7*).

The substantial disagreement among reviewers of the same proposals can be shown by a simple one-way analysis of variance for each group of reviewers (Table 4). About half of all the variance in ratings is seen to result from disagreement among reviewers of the same proposals. We replicated this one-way analysis of variance for the ten research programs studied in the first phase. In each of these programs we found that reviewer disagreement accounted for the largest share of the total variance in reviewer ratings. The within-proposal variance accounted for 35 to 63 percent of the total variance in the ten programs. Contrary to expectation, there was no less consensus in the social science fields of anthropology and economics than there was in the natural sciences (*8*).

Another way of conceptualizing the relatively low consensus among reviewers of the same proposal is to think of placing all the NSF reviews of the 50 chemical dynamics proposals in a hat and drawing two out at random. If we did this a large number of times we would find, on average, an expected absolute difference in the ratings of 9.78 points.

Now, if we placed all the reviews of a single proposal in the hat and drew out two, we would find, on average after multiple trials, an expected absolute difference of 8.45 points (*9*).

Conclusions

We have shown that the reversals observed in the COSPUP experiment can be explained by the substantial disagreement among reviewers of the same proposal. If getting an NSF grant were an entirely random process, we would have found a reversal rate approximating 50 percent. If, instead of conducting an independent evaluation of the proposals, we had simply flipped a coin to determine which of the 50 proposals evaluated by the NSF would be funded, we would have obtained a 50 percent reversal rate. The difference between what we would expect from the coin flip and what we observe with the data can be viewed as a measure of what we "buy" from the peer review process. Since the reversal rate is about 25 percent, we may conclude that the fate of a particular grant application is roughly half determined by the characteristics of the proposal and the principal investigator, and about half by apparently random elements which might be characterized as the "luck of the reviewer draw."

Although we conclude that the funding of a specific proposal submitted to the NSF is to a significant extent dependent on the applicant's luck in the program director's choice of reviewers, this should not be interpreted as meaning either that the entire process is random or that each individual reviewer is evaluating the proposal in a random way. In order to clarify the way in which the luck of the draw works, we must look at the sources of reviewer disagreement.

Some of the observed differences among scores given to the same proposal by different reviewers is undoubtedly an artifact of what anthropologists refer to as intersubjectivity. That is, there may be two reviewers who translate their substantively identical opinions differently; reviewer A's opinion is expressed as an "excellent" and reviewer B's as a "very good."

The great bulk of reviewer disagreement observed is probably a result of real and legitimate differences of opinion among experts about what good science is or should be. This became evident from the qualitative comments reviewers made both on the proposals studied for the COSPUP experiment and on those studied in the first phase of the peer review study. Contrary to a widely held belief that science is characterized by wide agreement about what is good work, who is doing good work, and what are promising lines of inquiry, our research both in this and other studies in the sociology of science indicates that concerning work currently in process there is substantial disagreement in all scientific fields (*10*).

As long as substantial reviewer disagreement, whatever its source, exists the fate of a particular proposal will depend heavily upon which reviewers happen to be selected. The element of chance would, of course, be reduced by increasing the number of reviewers; the larger the sample of reviewers the less likely it is that the sample mean will differ significantly from the population mean. It remains unclear what types of disagreement would obtain if we examined other forms of peer review, such as the study section method used at the National Institutes of Health. If we found less reviewer disagreement in that context, would that indicate that study sections are a method for achieving intellectual consensus on the relative merits of research proposals, or would it simply reflect "artificial" consensus resulting from the influence of nonintellectual forces that are part of group dynamics? Our data cannot speak to this question, since we did not examine peer review in the form used at NIH.

We must begin to question whether a system in which funding decisions depend to a significant degree on chance is the most rational one (*11*). Here we will conclude with two observations. First, given the importance of chance in the current process, clearly the more proposals a researcher submits the higher the probability of being funded. In fact, eminent scientists may be more likely to be funded than less well-known ones not because their probability of success is greater for each submitted proposal, but because they submit many proposals and are not deterred by an individual rejection. Second, the primary way in which the effect of chance might be reduced might be to give more weight to criteria for which there would be greater agreement than there is on the proposal. For example, it might be easier for scientists to agree upon the value of recently completed work than upon the value of a proposed piece of work.

Several important questions arise from the finding that there is a substantial random element in who gets an NSF grant. What degree of precision should we expect from the peer review system? Is it not healthy for science to have substantial disagreement among scientists who evaluate proposals, rather than a single, agreed-upon dogma? At what point does disagreement become dysfunctional for the development of science? A distinction must be made between the effect of randomness in the peer review system on individual applicants and the effect on science itself. Plainly, the random element can be frustrating and debilitating for individual scientists trying to obtain financial support for their work, but it may have little effect on the rate of development of science as a whole. One clear disadvantage for science of the current peer review system is that it compels even our most talented scientists to spend substantial amounts of time and energy writing proposals, time and energy that might be more fruitfully spent doing research.

References and Notes

1. S. Cole, L. Rubin, J. R. Cole, *Peer Review in the National Science Foundation: Phase I of a Study* (National Academy of Sciences, Washington, D.C., 1978). Peer review in the form used at the National Institutes of Health (NIH) was not studied. The NIH study section form of peer review differs significantly from the form of peer review at NSF. The results reported in this article apply only to the basic science programs at NSF and may not be generalizable to any other form of peer review.
2. The report of the first phase of that study (*1*) includes detailed descriptions of the central role of the NSF program directors in the decision-making process, the role of peer review panels, the influence of budget size on decisions, the effects of "self-selection" processes on peer review ratings, an analysis of problematic and unproblematic cases, and the correlates of peer review ratings and funding decisions. See also S. Cole, L. Rubin, J. R. Cole, *Sci. Am.* **237**, 34 (October 1977).
3. The 150 proposals were reviewed under two different conditions in the COSPUP experiment, only one of which is reported on in this article. Under one condition (the one dealt with here), the COSPUP reviewers were sent the proposal exactly as it was received by NSF, with the name and institution of the principal investigator listed on the title page, and were given exactly the same evaluation criteria to follow as were NSF reviewers. The participants in the experiment were told that their opinions would not influence the funding decision, which had already been made.

   Under the second condition, the proposal was altered so as to try to hide the identity of the principal investigator. A report on this part of the experiment can be found in J. R. Cole and S. Cole, *Peer Review in the National Science Foundation: Phase II of a Study* (National Academy of Sciences, Washington, D.C., 1981). The principal result of this part of the experiment was that it proved almost impossible to remove all evidences of the applicants' identity without destroying the integrity of proposals; the reversal rates for anonymous proposals were somewhat higher than for identified proposals; and the reasons for reversals were the same. A report on results of this part of the experiment is in preparation.

   The reviewer selectors received a copy of the proposal that did not include the name of the principal investigator and from which all references to the principal investigator's prior work had been deleted. It was therefore possible for the reviewer selector to name the principal investigator as a possible reviewer. In these cases, of course, the principal investigator's name was removed from the list of reviewers. We also removed from the list anyone who was

at the institution of the principal investigator and all who had reviewed the proposals for NSF.

4. The size of the pool (N) can be estimated as follows. This approximation assumes that each COSPUP selector makes six equiprobable choices from the pool of N. Suppose that individual A had been selected by NSF. Then P (A will be selected by COSPUP selector) = $6/N$. Let $p = 6/N$. Then P (no overlap) = $(1 - p)^2$; P (one overlap) = $2p\ (1 - p)$; P (double overlap) = p^2. The proportions $(1 - p)^2$: $2p\ (1 - p)$: p^2 correspond closely to the proportions observed, when $p = 0.1$. This suggests that $N \doteq 60$. The approximation is not perfect, of course. The field of economics produces surprisingly many double overlaps.
5. NSF may decide to fund a piece of the proposed scientific work, reducing the amount of the grant accordingly. Here we do not differentiate such a grant from a full grant.
6. We classified reversals within quintiles as follows: (i) The proposals were grouped into quintiles based on COSPUP rank, the first (best) quintile containing proposals with ranks 1, 2, . . . , 10. (ii) A proposal was counted as reversed if it was in the upper 25 by one set of ratings and in the lower 25 by the other set. (iii) Where there were ties in the mean ratings crossing quintile boundaries, proposals were apportioned among the categories involved. This rule results in noninteger numbers of reversals.
7. The relatively small differences in proposal means may result from processes of self-selection; that is, perhaps only relatively good scientists apply to NSF. This self-selection may result in attenuation of variance in reviewer ratings.
8. In all situations there were about the same number of proposals and reviewers per proposal, so the R^2 values are comparable [J. R. Cole and S. Cole, *Nature (London)* **279**, 575 (1979)].
9. If one takes two independent observations from a normal population with standard deviation σ, the expected absolute difference is $2\ \sigma/\sqrt{\pi}$. This statement is a reasonable approximation even when the population values do not follow a normal distribution. The numbers given in the text reflect a reviewer standard deviation of 7.49 and a proposal standard deviation of 4.36. Note that $(2/\sqrt{\pi})\ 7.49 = 8.45$ and $(2/\sqrt{\pi})\ (4.36^2 + 7.49^2)^{1/2} = 9.78$. The value $(4.36^2 + 7.49^2)^{1/2}$ reflects the fact that a randomly selected review incorporates both the proposal standard deviation and the reviewer standard deviation.
10. S. Cole, J. R. Cole, L. Dietrich, in *Toward a Metric of Science: The Advent of Science Indicators*, Y. Elkana, J. Lederberg, R. K. Merton, A. Thackray, H. Zuckerman, Eds. (Wiley, New York, 1978), p. 162.
11. A paper on how the system could be changed and how alternative systems might be more reasonable is in preparation.
12. The research was conducted under a contract between NSF and COSPUP. The statistical model was designed by J. Kiefer. We thank L. Cronbach for help in applying the model and for reviewing the results, the many members of COSPUP for their many useful suggestions in the design and analysis of the experimental material, and I. M. Singer for his support, without which the experiment could not have been carried out.

AAAS–Newcomb Cleveland Prize
To Be Awarded for an Article or a Report Published in *Science*

The AAAS–Newcomb Cleveland Prize is awarded annually to the author of an outstanding paper published in *Science* from August through July. This competition year starts with the 7 August 1981 issue of *Science* and ends with that of 30 July 1982. The value of the prize is $5000; the winner also receives a bronze medal.

Reports and Articles that include original research data, theories, or syntheses and are fundamental contributions to basic knowledge or technical achievements of far-reaching consequence are eligible for consideration for the prize. The paper must be a first-time publication of the author's own work. Reference to pertinent earlier work by the author may be included to give perspective.

Throughout the year, readers are invited to nominate papers appearing in the Reports or Articles sections. Nominations must be typed, and the following information provided: the title of the paper, issue in which it was published, author's name, and a brief statement of justification for nomination. Nominations should be submitted to AAAS–Newcomb Cleveland Prize, AAAS, 1515 Massachusetts Avenue, NW, Washington, D.C. 20005. Final selection will rest with a panel of distinguished scientists appointed by the Board of Directors.

The award will be presented at a session of the annual meeting. In cases of multiple authorship, the prize will be divided equally between or among the authors.

Deadline for nominations: postmarked 16 August 1982

Nomination Form
AAAS–Newcomb Cleveland Prize

AUTHOR: ____________________

TITLE: ____________________

____

DATE PUBLISHED: ____________________

JUSTIFICATION: ____________________

____

____

____

____

____

____

____